НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

ФАКУЛЬТЕТ ПРИКЛАДНОЇ МАТЕМАТИКИ

Кафедра системного програмування і спеціалізованих комп‘ютерних систем

«До захисту допущено»

Завідувач кафедри

_____ Тарасенко В.П.
(підпис) (ініціали, прізвище)

“___” червня 2019 р.

Дипломний проект

на здобуття ступеня бакалавра

з напряму підготовки **6.050102 «Комп'ютерна інженерія»**

на тему: “ Комп'ютерна система реєстрації новин на веб-ресурсах”.

Виконав: студент IV курсу, групи КВ-51

Скрипник Владислав Сергійович _____
(підпис)

Керівник, проф. каф. СПіСКС, д.т.н. проц. Терейковський І.А. _____
(підпис)

Консультант з нормоконтролю, доц.каф.СПіСКС, к.т.н. Клятченко Я.М. _____
(підпис)

Рецензент_____ _____
(посада, науковий ступінь, вчене звання, науковий ступінь, прізвище та ініціали) (підпис)

Засвідчую, що у цьому дипломному проекті немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ – 2019 року

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

ФАКУЛЬТЕТ ПРИКЛАДНОЇ МАТЕМАТИКИ

Кафедра системного програмування і спеціалізованих комп‘ютерних систем

Рівень вищої освіти – перший (бакалаврський)

Напрям підготовки 6.050102 «Комп’ютерна інженерія»

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Тарасенко В.П.
(підпис) (ініціали, прізвище)
«___»червня 2019 р.

ЗАВДАННЯ
на дипломний проект студента
Скрипник Владислава Сергійовича

1. Тема проекту: “Комп'ютерна система реєстрації новин на веб-ресурсах”.

Керівник проекту: проф. каф. СПіСКС, д.т.н. проц. Терейковський І.А.,

затверджені наказом по університету від «22»травня 2019 р. №1330-С

2. Термін подання студентом проекту _____

3. Вихідні дані до проекту:див. технічне завдання.

4. Зміст пояснювальної записки:аналіз існуючих рішень та обґрунтування тем, проблематика розпізнавання деструктивного контенту, розробка інструментальних засобів комп’ютерної системи реєстрації новин на веб-ресурсах, особливості та переваги розробленої системи.

5. Перелік графічного матеріалу (із зазначенням обов’язкових креслеників, плакатів, презентацій тощо): алгоритм захоплення новин з веб-ресурсів, алгоритм реєстрації новин в базі даних, алгоритм інтерпретації тексту новин

системою, алгоритм роботи комп'ютерної системи реєстрації новин на веб-ресурсах, структура комп'ютерної системи реєстрації новин на веб-ресурсах.

6. Консультанти розділів проекту*

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Нормоконтроль	Клятченко Я.М., доц. каф. СПіСКС, к.т.н.		

7. Дата видачі завдання_____

Календарний план

№ з/п	Назва етапів роботи та питань, які мають бути розроблені відповідно до завдання	Термін виконання
1.	Видача завдання на дипломне проектування	04.11.2018
2.	Вивчення літератури за тематикою роботи	16.11.2018
3.	Розроблення та узгодження технічного завдання	17.01.2019
4.	Розроблення структури системи	09.02.2019
5.	Розроблення алгоритмів системи	04.03.2019
6.	Аналіз системи	04.04.2019
7.	Підготовка матеріалів текстової частини проекту	29.04.2019
8.	Підготовка матеріалів графічної частини проекту	18.05.2019
9.	Оформлення технічної документації проекту	27.05.2019

Студент _____ ___Скрипник В.С.____
(підпис) (ініціали, прізвище)

Керівник проекту _____ __ Терейковський І.А._
(підпис) (ініціали, прізвище)

*Консультантом не може бути зазначено керівника дипломного проекту.

АНОТАЦІЯ

Кваліфікаційна робота включає пояснювальну записку (51 с., 13 рис., 2 додатки).

Бакалаврський проект призначено розробці комп'ютерної системи реєстрації новин на веб-ресурсах, яка дозволяє визначити чи містить блок новин веб-ресурса деструктивний контент.

Комп’ютерна система реєстрації новин на веб-ресурсах дозволяє: знайти і проаналізувати блок новин на веб-ресурсі за посиланням; здійснити аналізу, як тексту сторінки в цілому, так і його окремих блоків на наявність контенту деструктивного характеру; проводити повторення аналізу веб-ресурсу для випадків коли веб-ресурс генерує список новин при кожному відкритті сторінки. Передбачена можливість не тільки виявлення яскраво вираженого деструктивного контенту за рахунок слів-маркерів, але й винесення підозри про наявність прихованого деструктивного контенту за рахунок аналізу стиля написання новини.

В ході розробки:

- проведено аналіз існуючих рішень для боротьби з деструктивним контентом в Інтернеті;
- сформульовані вимоги до комп’ютерної системи реєстрації новин на веб-ресурсах;
- розглянута проблематика розпізнавання деструктивного контенту;
- розроблені алгоритми функціонування комп’ютерної системи реєстрації новин на веб-ресурсах.

Використання системи наведеної в цій роботі дозволяє користувачам оцінити наявність деструктивного контенту на веб-ресурсі перед його відвідуванням.

Ключові слова:

НЕЙРОННІ МЕРЕЖІ, МАШИННЕ НАВЧАННЯ, PYTHON, SCIKIT-LEARN, SCRAPY, URLLIB, NOSQL, MONGODB.

ABSTRACT

The diploma project includes an explanatory note (51 p., 13 fig., 2 appendices).

The Bachelor's project is designed to develop a computer system for registering news on web resources, which allows user to determine whether the web site`s news block contains destructive content or not.

The mobile application can perform next actions : find and analyze a news block on a web resource; analyze text of the page as a whole and its individual blocks for the presence of destructive content; repeat the analysis of a web resource for cases when a web resource generates a news list every time a page is opened. Also, it provides capability to not only to reveal a destructive content at the expense of word markers, but also to suspect that text contains hidden destructive content is at the expense of analyzing the style of writing news.

In the development process were resolved:

- analysis of existing solutions to combat destructive content on the Internet;
- formulation of requirements for computer system of registration of news;
- consideration of the the problem of recognition of destructive content;
- algorithms of functioning of the computer system of registration of news on web-resources.

Using the system provided in this work allows users to evaluate the presence of destructive content on a web site before visiting it.

Key words:

NEURAL NETWORKS, MACHINE LEARNING, PYTHON, SCIKIT-LEARN, SCRAPY, URLLIB, NOSQL, MONGODB.

Поз.	Формат	ПОЗНАЧЕННЯ	НАЙМЕНУВАННЯ	Кількість аркушів	№ прим.	Примітки
	А4	ІАЛЦ.045492.002 ТЗ	Комп'ютерна система	4		
			реєстрації новин на			
			веб-ресурсах			
			Технічне завдання			
	А4	ІАЛЦ.045492.003 ТП	Комп'ютерна система	2		
			реєстрації новин на			
			веб-ресурсах			
			Відомість технічного			
			проекту			
	А4	ІАЛЦ.045492.004 ПЗ	Комп'ютерна система	51		
			реєстрації новин на			
			веб-ресурсах			
			Пояснювальна записка			
	А4	ІАЛЦ.045492.005 Д1	Алгоритм захоплення	1		
			новин з веб-ресурсів.			
			Схема алгоритму.			
	А4	ІАЛЦ.045492.006 Д2	Алгоритм реєстрації	1		
			новин в базі даних.			
			Схема алгоритму.			

Змін.	Арк.	№ докум.	Підпис	Дата	ІАЛЦ.045492.001 ОА		
Розробив		Скрипник В.С.			Комп'ютерна система реєстрації новин на веб-ресурсах **Опис альбому**	Літ.	Аркуш 1 / Аркушів 2
Перевірив		Терейковський І. А.					
Консульт.						КПІ ім. Ігоря Сікорського, ФПМ КВ-51	
Н. контроль		Клятченко Я.М.					
Зав. каф.		Тарасенко В.П.					

Поз.	Формат	ПОЗНАЧЕННЯ	НАЙМЕНУВАННЯ	Кількість аркушів	№ прим.	Примітки
	А4	ІАЛЦ.045492.007 Д3	Алгоритм інтерпретації	1		
			тексту новин системою.			
			Схема алгоритму.			
	А4	ІАЛЦ.045492.008 Д4	Алгоритм реєстрації	1		
			комп’ютерної системи			
			реєстрації новин на			
			веб-ресурсах			
			Схема алгоритму.			
	А4	ІАЛЦ.045492.009 Д5	Структура комп’ютерної	1		
			Системи реєстрації новин			
			на веб-ресурсах.			
			Схема структурна.			
		Диск CD-ROM	Текст пояснювальної	1		
			записки.			
			Графічний матеріал			

ЗМІСТ

1. НАЙМЕНУВАННЯ ТА ГАЛУЗЬ РОЗРОБКИ

Назва розробки: «Комп'ютерна система реєстрації новин на веб-ресурсах».

Галузь застосування: користувач всесвітньої мережі, що бажають уникати веб-ресурсів, що містять контент деструктивного характеру.

2. ПІДСТАВА ДЛЯ РОЗРОБКИ

Підставою для розробки є завдання на виконання роботи першого (бакалаврського) рівня вищої освіти, затверджене кафедрою системного програмування і спеціалізованих комп’ютерних систем Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського».

3. МЕТА І ПРИЗНАЧЕННЯ РОБОТИ

Метою даного проекту є розробка комп’ютерної системи реєстрації новин на веб-ресурсах для проведення їх аналізу на предмет наявності в них контенту деструктивного характеру.

4. ДЖЕРЕЛА РОЗРОБКИ

Джерелом інформації є технічна та науково-технічна література, технічна документація, публікації в періодичних виданнях та електронні статті у мережі Інтернет.

5. ТЕХНІЧНІ ВИМОГИ

5.1. Вимоги до програмного продукту, що розробляється

- Наявність функції аналізування веб ресурса на предмет новин, отриманих з зовнішніх джерел, що містять деструктивний контент;
- Проведення аналізу не тільки по всьому тексту веб-сторінки, але й по окремим блокам;
- Зберігання інформації про проведений аналіз веб-сторінки в базу-даних.

5.2. Вимоги до апаратного забезпечення

- оперативна пам’ять: 1 Гб;
- частота процесора: 2 ГГЦ.

5.3. Вимоги до програмного та апаратного забезпечення користувача

- операційна система Linux;
- наявність доступу до мережі інтернет.

6. ЕТАПИ РОЗРОБКИ

№ з/п	Назва етапів виконання дипломного проекту	Термін виконання етапів
1.	Видача завдання на дипломне проектування	04.11.2018
2.	Вивчення літератури за тематикою роботи	16.11.2018
3.	Розроблення та узгодження технічного завдання	17.01.2019
4.	Розроблення структури системи	09.02.2019
5.	Розроблення алгоритмів системи	04.03.2019
6.	Аналіз системи	04.04.2019
7.	Підготовка матеріалів текстової частини проекту	29.04.2019
8.	Підготовка матеріалів графічної частини проекту	18.05.2019
9.	Оформлення технічної документації проекту	27.05.2019

Поз.	Формат	ПОЗНАЧЕННЯ	НАЙМЕНУВАННЯ	Кількість аркушів	№ прим.	Примітки
	А4	ІАЛЦ.045492.004 ПЗ	Комп'ютерна система	51		
			реєстрації новин на			
			веб-ресурсах			
			Пояснювальна записка			
	А4	ІАЛЦ.045492.005 Д1	Алгоритм захоплення	1		
			новин з веб-ресурсів.			
			Схема алгоритму.			
	А4	ІАЛЦ.045492.006 Д2	Алгоритм реєстрації	1		
			новин в базі даних.			
			Схема алгоритму.			
	А4	ІАЛЦ.045492.007 Д3	Алгоритм інтерпретації	1		
			тексту новин системою.			
			Схема алгоритму.			
	А4	ІАЛЦ.045492.008 Д4	Алгоритм реєстрації	1		
			комп’ютерної системи			
			реєстрації новин на			
			веб-ресурсах			
			Схема алгоритму.			

Змін.	Арк.	№ докум.	Підпис	Дата	ІАЛЦ.045492.003 ТП		
Розробив		Скрипник В.С.			Комп'ютерна система реєстрації новин на веб-ресурсах **Відомість технічного проекту**	Літ.	Аркуш 1 / Аркушів 2
Перевірив		Терейковський І. А.					
Консульт.						КПІ ім. Ігоря Сікорського, ФПМ КВ-51	
Н. контроль		Клятченко Я.М.					
Зав. каф.		Тарасенко В.П.					

Поз.	Формат	ПОЗНАЧЕННЯ	НАЙМЕНУВАННЯ	Кількість аркушів	№ прим.	Примітки
	А4	ІАЛЦ.045492.009 Д5	Структура комп’ютерної	1		
			Системи реєстрації новин			
			на веб-ресурсах.			
			Схема структурна.			
		Диск CD-ROM	Текст пояснювальної	1		
			записки.			
			Графічний матеріал			

Змін.	Арк.	№ докум.	Підпис	Дата	ДП.468300.003 ТП	Арк. 2

ЗМІСТ

Додаток 1. Копії графічних матеріалів

- ІАЛЦ.045492.005 Д1. Алгоритм захоплення новин з веб-ресурсів. Схема алгоритму.
- ІАЛЦ.045492.006 Д2. Алгоритм реєстрації новин в базі даних. Схема алгоритму.
- ІАЛЦ.045492.007 Д3. Алгоритм інтерпретації тексту новин системою. Схема алгоритму.
- ІАЛЦ.045492.008 Д4. Алгоритм роботи комп’ютерної системи реєстрації новин на веб-ресурсах. Схема алгоритму.
- ІАЛЦ.045492.009 Д5. Структкура комп’ютерної системи реєстрації новин на веб-ресурсах. Схема структурна

Додаток 2. Презентація

ПЕРЕЛІК СКОРОЧЕНЬ, УМОВНИХ ПОЗНАЧЕНЬ, ТЕРМІНІВ

API (Application Programming Interface) – прикладний програмний інтерфейс;

DDoS-атака, або DDoS (Distributed Denial of Service) – розподілена атака на відмову в обслуговуванні;

DOM (Document Object Model) – специфікація прикладного програмного інтерфейсу для роботи зі структурованими документами (як правило, документами XML);

Брандмауер або Файрвол (Firewall) – мережевий екран;

Веб-браузер (web browser) – програмне забезпечення для комп'ютера або іншого електронного пристрою, як правило, під'єднаного до Інтернету, що дає можливість користувачеві взаємодіяти з текстом, малюнками або іншою інформацією на гіпертекстовій веб-сторінці;

Веб-краулер (WEB crawler) – пошуковий робот;

Веб-ресурс (web resource) – ресурс, який розміщенно в просторі мережі Інтернет;

Веб-форми (WEB forms) – елемент веб-сторінки, який дає користувачам можливість вводити інформацію і відправляти її на сервер для подальшої обробки;

Геолокація (Geolocation) – ідентифікацію або оцінку реального географічного розташування об'єкта;

ЗМІ (Mass media) – засоби масової інформації;

Інтернет (Internet) – всесвітня система сполучених комп'ютерних мереж, що базуються на комплекті Інтернет-протоколів;

Капча (Completely Automated Public Turing test to tell Computers and Humans Apart) – повністю автоматизований публічний тест Тюринга для розрізнення комп'ютерів і людей;

Куки (Cookie) – інформація у вигляді текстових або бінарних даних, отриманих від веб-сайту на веб-сервері, яка зберігається у клієнта, тобто браузера, а потім відправлена на той самий сайт, якщо його буде повторно відвідано;

Майнер (Miner) – приховане програмне забезпечення на веб-сторінці, що використовує процесорні потужності клієнта для отримання вигоди в вигляді критовалюти;

МН – Машинне Навчання;

Нейроні Мережі (Neural Network) – обчислювальні системи, натхнені біологічними нейронними мережами, що складають мозок тварин;

Плагін (Plugin) – додаток, незалежно скомпільований програмний модуль, що динамічно підключається до основної програми, призначений для розширення або використання її можливостей;

Проксі-сервер, або проксі (Proxy) – сервер (комп'ютерна система або програма) в комп'ютерних мережах, що дозволяє клієнтам виконувати непрямі (через посередництво проксі-сервера) запити до мережевих сервісів;

Скрипт (Script) – сценаріїй, послідовність операцій, що зберігаються в зрозумілому людині виді, які виконується комп'ютером;

Трекер (Tracker) – приховане програмне забезпечення на веб-сторінці, що відслідковує активність користувача;

Шкідливе програмне забезпечення (Malware) – програмне забезпечення, яке перешкоджає роботі комп'ютера, збирає конфіденційну інформацію або отримує доступ до приватних комп'ютерних систем.

ВСТУП

На сьогодні існує велика кількість сайтів, що надають вибірку з декількох новин, що можуть бути актуальні для користувача. Які саме новини будуть показані базується на інформації про користувача – його геолокація, походження, мова спілкування, місце проживання, покупки в магазинах, історія веб-браузера, інформація користувачів із його кола спілкування, тощо. Звісно така система є дуже корисною, і в більшості випадків надає користувачу новини в яких він зацікавлений, проте алгоритм пошуку таких новин не є ідеальним і серед результатів можуть бути новини, що не задовольняють критерії користувача. Також більшість таких сайтів беруть посилання на новини з відкритих ресурсів і серед результатів їх пошуку може бути контент, що несе деструктивний характер.

Головною проблемою сайтів, що показують своїм користувачам вибірки новин є те, що результат вибірки генерується враховуючи дані користувача та актуальність новин і через це користувачі можуть зіткнутися з великою кількістю негативного та деструктивного контенту серед представлених новин. Такі новини можуть бути просто не приємні користувачу, а в найгірших випадків вони можуть нести призиви до екстремізму, фаворитизм до тероризму та багато іншого контенту деструктивного характеру. Особливу небезпеку це несе для молодих людей та дітей – в кращому випадку це просто псує настрій та призводить до накоплення стресу, а в найгіршому випадку деструктивні новини можуть впливати на підсвідомість дітей та приводити до самолінчування, самогубств або участі в організованій злочинності або тероризму.

Даний дипломний проект націлений на розробку системи, що може захоплювати новини на таких сайтах, фіксувати їх в базі даних, аналізувати на наявність деструктивного контенту та видавати висновок

про безпечність користування таким веб-ресурсом. Головною метою цієї системи є створення безпечного середовища в інтернеті, що зменшує кількість стресу, що переживають користувачі інтернету, а також захищає ментально не захищених користувачів від сайтів, що можуть містити деструктивний контент.

Розроблена система надає користувачам можливість не тільки перевірити веб-ресурс перед відвідуванням, що не тільки напряму допомагає користувачу проаналізувати кількість деструктивного контенту наданого даним ресурсом, але й спонукає розробників та власників цих ресурсів розроблювати кращі алгоритми та фільтри пошуку новин для їх відвідувачів.

Метою дипломної роботи є створити комп'ютерну систему реєстрації новин на веб-ресурсах, що дозволить зменшити ментальну втому користувачів інтернету за рахунок виявлення сайтів, блок з новинами яких містить контент деструктивного характеру.

Об'єктом дослідження є процес ідентифікації вмісту прихованого та явного деструктивного контенту в текстах новин за допомогою згорткових нейронних мереж.

Предмет дослідження – нейромережеві моделі та методи якісно-кількісного аналізу тексту.

Структурно дипломна робота складається із вступу, чотирьох розділів, висновків, списку використаної літератури та двох додатків.

У першому розділі проведено аналіз задачі захисту користувачів всесвітньої мережі інтернет від контенту деструктивного характеру. Для цього охарактеризовано проблему розробки вказаних засобів та проведено огляд відповідних технологій. В результаті проведено обґрунтування задач дипломної роботи.

У другому розділі здійснений аналіз методів розпізнавання контенту деструктивного характеру в тексті веб-сторінки, що дозволило побудувати відповідні алгоритми та розробити математичне забезпечення.

У третьому розділі дипломної роботи наведена розробка програмного забезпечення комп’ютерної системи реєстрації новин на веб-ресурсах. Розробка складається із побудови алгоритмів функціонування основних модулів системи та створення її структури.

У четвертому розділі проведений аналіз особливостей та переваг розробленої системи реєстрації новин на веб-ресурсах.

1. АНАЛІЗ ІСНУЮЧИХ РІШЕНЬ ТА ОБҐРУНТУВАННЯ ТЕМИ

1.1. Загальна характеристика проблеми

Для користувачів глобальної інформаційної мережі Інтернет в останній час важливості набирає питання наявності контенту деструктивного характеру, що складається з тролінгу, віртуального насилля, агресії, ворожості, тощо. Ця проблема не є нова, ще до появи Інтернету існували індивіди, що наповнювали життя людей негативним та деструктивним контентом для втілення власних цілей – будь то самозадоволення, або навіть підштовхування до державних переворотів та силового захоплення влади.

Зростання деструктивних настроїв у суспільстві свідчить про незадоволеність населення рівнем та якістю життя, та може приводити до дестабілізації. Один з важливих факторів в такому процесі відіграють інформаційні ресурси, зокрема Інтернет ресурси, що представляють новини. Особливу небезпеку становлять Інтернет ресурси, що використовують стороні ресурси для побудови своєї стрічки новин використовуючи алгоритми машинного навчання. Такі ресурси не тільки не можуть відповідати за контент, що з’являється в їхній стрічці новин, але й допускають появу новин, що містять приховані заклики до екстремізму, а також збільшує популярність таких новин шляхом збільшення кількості відвідувачів Інтернет сторінки новини, тим самим збільшуючи ймовірність появи такої новини у стрічках інших користувачів, які використовують відвідуваність Інтернет сторінки новини, як один з критеріїв, що визначає релевантність новини для користувача.

Соціологічні та психологічні науки розрізняють два основних види аналізу документів – якісний (традиційний) і кількісний, який за

міжнародною класифікацією називають контент-аналізом [1]. Контент-аналіз полягає в переведені масиву інформації в кількісні показники. Саме цей вид аналізу найкраще підходить для використання в нашій комп'ютерній системі реєстрації інформації, так як дозволяє дослідити повідомлення на деструктивний контент не тільки за наявністю слів-маркерів, що вказують на деструктивний характер повідомлення, але й проаналізувавши частоту використання інших слів повідомлення.

Для зручності можна розбити деструктивний контент на категорії:

- насилля – контент, що містить заклики до насильства, або виражає позитивне ставлення до випадків насилля;
- ворожість – контент, що містить вираження ворожості по відношенні до іншої соціальної групи, національності, тощо;
- агресія – контент, що містить вираження агресії по відношенню до читача, цей контент має на меті викликати негативну реакцію з боку читача.

Деструктивний контент в будь-якій його формі призводить до великої кількості негативний наслідків: погіршенню настрою, збільшення ймовірності появи депресії, тощо [2]. Така ситуація потребує уваги зі сторони суспільства і не може бути розв'язана за рахунок природного ходу речей.

В мережі Інтернет існують сайти, що надають своїм відвідувачам можливість дізнатися актуальні для них новини з різних джерел в одному місці, наприклад ukr.net (рис. 1.1) [3], google (рис. 1.2) [4], yandex (рис. 1.3) [5], тощо. Зазвичай такі сайти не мають власних команд ЗМІ (засоби масової інформації) і тому використовують інформацію з відкритих джерел. Звісно така інформація може бути корисною користувачу, проте можливі випадки коли серед результатів вибірки з'являються матеріали

деструктивного характеру. Така ситуація є навіть більш небезпечною ніж здається на перший погляд. Алгоритми пошуку новин зазвичай базуються на ряді факторів, серед яких одним з основних є популярність, тобто відвідуваність, сторінки новини. Таким чином при появі новини в стрічці користувача є ймовірність того, що він перейде за посиланням і перегляне повну версію новини, тим самим збільшивши її рейтинг і таким чином збільшивши ймовірність того, що цю новину буде показано іншим користувачам. Звісно автори деструктивних текстів про це добре знають і тому при оформлені новини з деструктивним контентом вони зазвичай створюють заголовок, який має зацікавити користувача відкрити посилання. Такий прийом в Інтернеті отримав назву «клік-байтинг», що в перекладі з англійської означає «наживка на кліки» [6]. Цей прийом отримав значне поширення, особливо серед авторів новин з деструктивним контентом.

Рисунок 1.1 – Головна сторінка ukr.net надає відвідувачам вибірку з новин на відкритих ресурсах по категоріям «Головне», «Політика», «Економіка», тощо

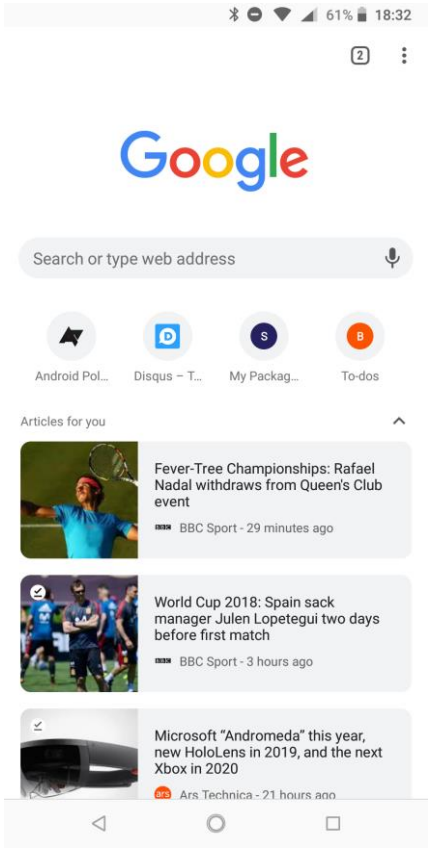

Рисунок 1.2 – Мобільний додаток Chrome від Google в розділі «Articles for you» надає користувачам персоналізовані вибірки з зовнішніх джерел

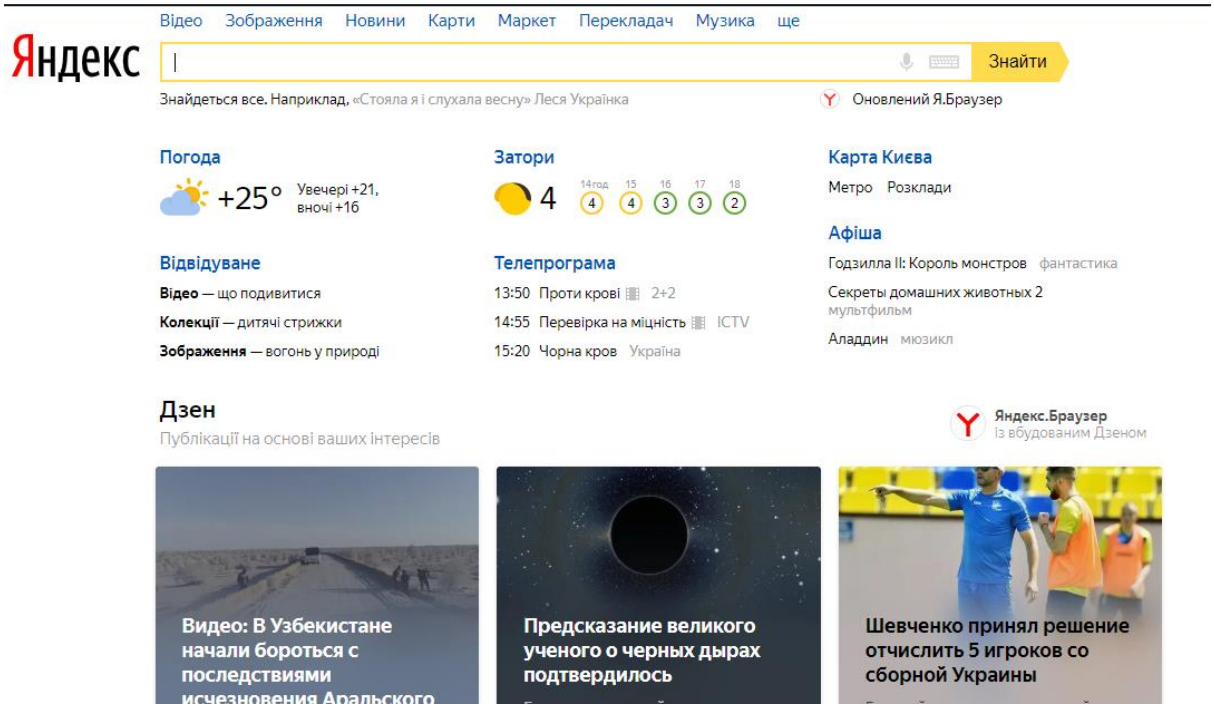

Рисунок 1.3 – Головна сторінка Yandex в розділі «Дзен» надає користувачам персоналізовані вибірки з зовнішніх джерел

Окрім новин, що несуть деструктивний характер, такі вибірки можуть нести й набагато гірший контекст, а саме прихований деструктивний характер. Такі новини, зазвичай не наносять шкоди відразу, а повільно накопичуються в свідомості читача і призводять до набагато гірших наслідків. Автори таких новин, приховують свої наміри в контексті тексту і тим самим не дають читачу одразу зрозуміти, що він піддався негативному впливу. Такі новини набагато складніше виявити, ніж новини з явним деструктивним контекстом, так як зазвичай вони містять кількість слів-маркерів, що не перебільшують кількість слів-маркерів у звичайних новинах. Це призводить до неможливості виявити такі новини просто порахувавши частоту зустрічі слів-маркерів у тексті. Ця ситуація потребує більш глибокого аналізу, що не можливо провести звичайним способом. В цій ситуації можуть допомогти нейронні мережі, які можна навчити визначати деструктивний контекст навіть за недостатньої кількості слів-маркерів в повідомленні. Звісно, це породжує нову проблему – підготовки інформації для навчання нейронної мережі, що потребує не тільки частотної характеристики тексту, але й більш чіткого аналізу, що зможе відрізнити нормальний текст від тесту з прихованим деструктивним контентом [6].

Новини, що несуть в собі прихований деструктивний контент є небезпечними для людей всіх вікових категорій. На відміну від новин з яскраво вираженим деструктивним контентом, новини що його приховують важко помітити навіть людям з багатим життєвим досвідом. При цьому, хоча й одноразова зустріч з новиною такого типу і не несе небезпеки ментальному здоров'ю людини, проте, якщо новини такого типу будуть зустрічатися більше ніж один раз в день, це може призвести до накоплення ментальної втоми і люди можуть більш легко піддаватись

маніпуляції. Тому, це є дуже важливо уникати такі новини, а для цього їх необхідно визначити.

За умови існування програми, що зможе проаналізувати новини, які представляє в своїх вибірках інтернет-ресурс, люди зможуть аналізувати сайти які вони відвідують і за необхідності знаходити альтернативу. Ця програма не тільки захистить людей від деструктивного контенту – явного і прихованого, але й спонукатиме власників інтернет-ресурсів розробляти більш досконалі алгоритми та фільтри для пошуку релевантних новин на відкритих інтернет джерелах.

1.2. Аналіз існуючих рішень для реєстрацій новин на веб-ресурсах

На сьогодні не існує готового рішення даної проблеми, проте існують засоби, що виконують схожі функції. До цих засобів належать:

- Блокувальниками реклами (рис. 1.4), які дозволяють відключити рекламу на інтернет-ресурсі, якщо користувач вважає, що вона несе негативний та/або деструктивний характер [7];
- Брандмауери (мережеві екрани) (рис. 1.5), які дозволяють користувачу заблокувати доступ до інтернет-ресурсу, якщо він вважає, що він несе негативний та/або деструктивний контент [8].

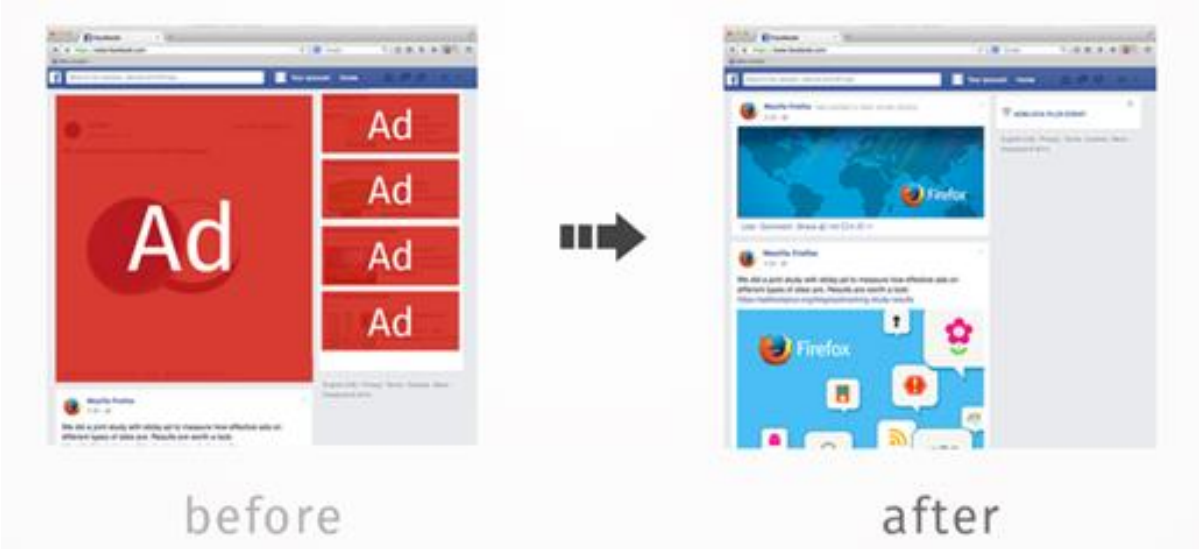

Рисунок 1.4 – Блокувальник реклами Adblock блокує деструктивну рекламу на інтернет-ресурсі

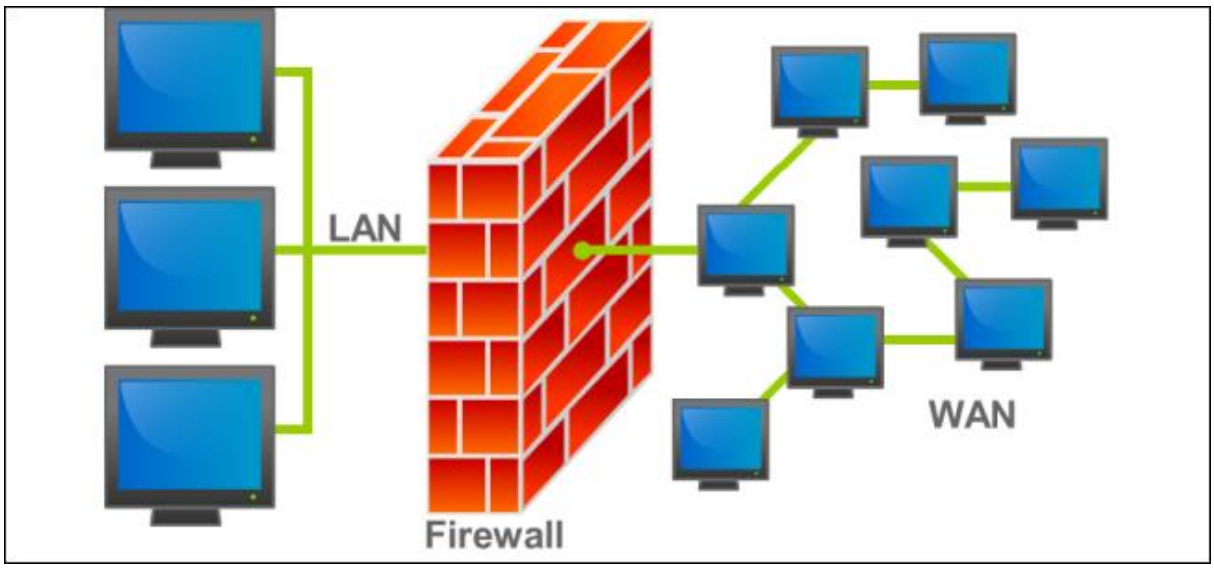

Рисунок 1.5 – Брандмауер (Firewall) фільтрує Інтернет трафік що надходить до користувача, захищаючи його від потенційно деструктивного контенту

Блокувальники реклами виникли через надлишкову кількість реклами на інтернет-ресурсах [7]. З початку вони несли негативний характер і більшість користувачів недолюблювали їх, проте з часом власники інтернет-ресурсів в погоні за наживою почали добавляти більше і більше рекламних секцій, часто сумнівного погодження. Це не тільки зменшило кількість корисної інформації на сторінки, але й принесло нову

загрозу, адже серед рекламних оголошень все частіше почали зустрічатись такі, що могли завдавати шкоди ментальному здоров'ю людей. Однією з небезпек, яка породжена зростаючою кількістю рекламних оголошень є те, що деякі з них можуть нести негативний, або деструктивний характер. Для боротьби з цим користувачі почали використовувати блокувальними реклами. Однією з функцією цих блокувальників є те, що користувачі самі можуть обирати блокувати рекламу на цьому ресурсі, чи ні. Найбільш сумлінні користувачі блокують рекламу тільки на тих ресурсах, де існує небезпека погіршити собі настрій через рекламу негативного або деструктивного характеру. Звісно є й такі користувачі, що настільки втомились від рекламних оголошень, більшість з яких несуть в собі деструктивну загрозу, що почали блокувати рекламу на всіх сайтах, а власників інтернет-ресурсів, які їм подобаються, підтримують за допомогою платних підписок та пожертв коштів [7].

До блокувальників реклами належать:

- uBlock Origin – блокує рекламу, трекери та шкідливе програмне забезпечення (malware) [9];
- Privacy Badger – відслідковує та за необхідністю блокує зовнішні домени на сторінки, що вбудовують зображення, скрипти та рекламні повідомлення на сторінці [10];
- Ghostery – відслідковує трекери та рекламні повідомлення на сторінці, надає інформацію про компанії яким належать ці треки та рекламні повідомлення та дає можливість заблокувати їх [11];
- AdGuard – блокує рекламу, трекери, шкідливе програмне забезпечення та майнери [12];
- AdBlock – блокує рекламу, трекери та шкідливе програмне забезпечення (malware) [13];

Звісно блокувачі реклами не гарантують повноцінний захист від деструктивного контенту, проте вони можуть зменшити загрузку на ментальне здоров'я людей, тим самим зменшити ймовірність того, що користувачі постраждають від новин деструктивного характеру. Утім, цього все ще не достатньо, щоб повноцінно захистити користувачів інтернет-ресурсів.

Інший засіб, яким користувачі можуть захистити себе від негативного та деструктивного контенту, це брандмауер також відомий як міжмережевий екран, мережевий екран, або файрвол [8]. Брандмаер дозволяє допускати, або відмовляти трафік, що надходять до користувача, згідно встановленого набору правил та інших критеріїв. Він може набувати вигляду окремого приладу (наприклад, роутер), або програмного забезпечення, що встановлено на персональний комп'ютер. В залежності від з'єднань, що відслідковуються, брандмауери розділяють на:

- проста фільтрація (stateless) – не відслідковують поточні з'єднання, а виконують фільтрування виключно на основі статичних правил [8];
- фільтрація з урахуванням контексту (stateful) – відслідковують поточні з'єднання та пропускають тільки такі пакети, які задовольняють алгоритми роботи відповідного протоколу чи програм. Такий тип брандмауерів дозволяє ефективніше протидіяти DDoS-атакам та вразливості деяких протоколів і мереж [8].

Так як в нашому випадку ціллю використання брандмауера є протидія контенту деструктивного характеру, то в нашому випадку найкраще підходить проста фільтрація. Таким чином користувач може створити «чорний список» сайтів та інтернет-ресурсів, матеріали яких він вважає небезпечними і передати цей список своєму брандмауеру. При

наступній спробі доступу до сайту з чорного списку користувач отримає повідомлення, що такий сайт не знайдено і тим самим не підпаде під вплив деструктивного та негативного контенту, що несе даний ресурс [8].

Очевидно, що хоча брандмауер захистить користувача від відвідувань небезпечних сайтів, проте користувачі все ще не будуть захищені від цитувань цих інтернет-ресурсів на інших сайтах. Звісно користувач буде захищений від повної версії небезпечного матералу, проте він може зазнати шкоди від інформації в заголовку, що буде представлено в вибірках новин на інших сайтах.

Також, обидва розглянуті засоби запобігання новин, які містять контент деструктивного характеру при відвідуванні інтернет-ресурсів, що представляють вибірки з відкритих джерел та інтернет-ЗМІ, мають значний недолік, який полягає в тому, що користувач повинен відвідати хоча б раз інтернет-ресурс щоб дізнатися чи містить він деструктивний контент. І хоча цей недолік несе небезпеку тільки при першому відвідувані сайту, проте він може набути критичної важливості у випадку коли користувач часто відвідує нові сайти.

1.3. Формулювання задач дипломної роботи

В результаті проведеного дослідження можна сформулювати такі вимоги до комп’ютерної системи реєстрації новин на веб-ресурсах:

- система повинна вміти аналізувати не тільки витяги з новин, що знаходяться на сторінці веб-ресурсу, але й тексти повної версії новин, які знаходяться за посиланнями розміщеними в блоці новин;
- система повинна аналізувати текст на наявність не тільки яскраво вираженого деструктивного контенту, але й прихованого;

– система повинна бути орієнтована на використання в комп’ютерних системах, що функціонують під управлінням сучасних операційних систем.

Крім того, проведений аналіз задачі розробки системи реєстрації новин на веб-ресурсах дозволяє виділити такі етапи її вирішення:

1. Побудова архітектури системи реєстрації новин на веб-ресурсах.
2. Розробка структури комп’ютерної системи.
3. Розробка інформаційного забезпечення.
4. Розробка програмного забезпечення.
5. Експериментальне підтвердження отриманих результатів.

Комп’ютерна система реєстрації новин на веб-ресурсах повина коректно обробляти інформацію веб-ресурсу. Це включає не тільки коректний аналіз новин на наявність в них контенту деструктивного характеру, але й коректне отримання тексту сторінки від веб-ресурсу.

2. ПРОБЛЕМАТИКА РОЗПІЗНАВАННЯ ДЕСТРУКТИВНОГО КОНТЕНТУ

2.1. Методи збору даних на веб-ресурсах

Методи збору даних на веб-ресурсах можна поділити на два основних типа:

1. браузерні – які потребують веб-браузер для своєї роботи і отримують дані інтернет сторінки від браузера, наприклад плагіни (plugins) [14];
2. дескотпні – які не потребують веб-браузера для своєї роботи і отримують дані інтернет сторінки напряму від веб-сервера через стандартні засоби системи, наприклад веб-краулери (web-crawlers) [15].

Звісно, на перший погляд перший тип має значну перевагу перед другим, так як може напряму функціонувати з веб-сторінкою, яку переглядає користувач, що надає можливість втілити більш функціональну систему, яка зможе одразу редагувати контент, що відображається браузером. Проте на справді збір даних першим способом має значну ваду – сильну залежність від браузера. Програми, що використовують методи збору даних першого типу сильно залежать від виробника веб-браузера. Це призводить не тільки до того, що програма може бути використана тільки в одному браузері, під який вона була розроблена, але й до того, що програма може перестати працювати через зміни в API (прикладний програмний інтерфейс) браузера його виробником [16].

Другий тип хоча й змушує користувача запускати окрему програму для збору даних на веб-ресурсі, проте користувач може не хвилюватися про те які версію програми він використовує, адже вони будуть

працювати, навіть після оновлення веб-браузера чи встановлені іншого веб-браузера.

Для розробки комп'ютерної системи реєстрації новин було обрано метод збору даних другого типу, а саме веб-краулер Scrapy [17].

Scrapy має відкритий вихідний код і дуже простий алгоритм роботи, тим самим він є ідеальним вибором для отримання тексту веб-сторінок, що будуть перевірятися на наявність деструктивного контенту в них. Для початку роботи зі Scrapy потрібно створити проект командою «scrapy startproject project», виконання цієї команди створить стандартний проект Scrapy (рис. 2.1).

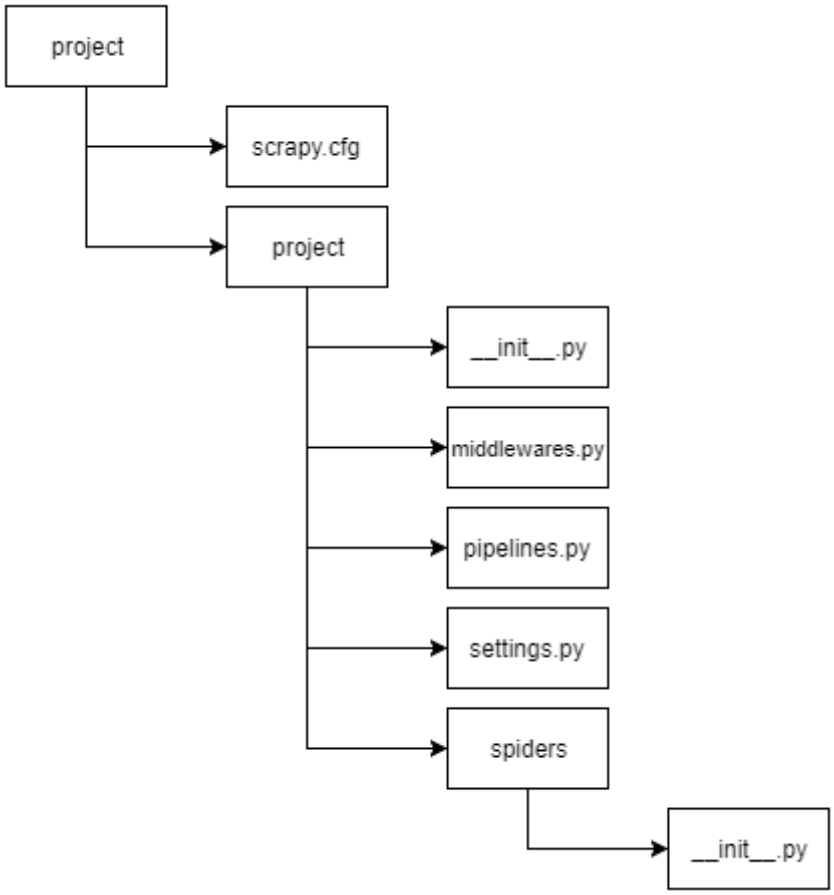

Рисунок 2.1 – Стандартна структура проекту Scrapy

Стандартний проект Scrapy [17] складається з наступних компонентів:

- scrappy.cfg – файл конфігурації розгортання (deploy) проекту;
- project – директорія-модуль Python проекту;
- project/items.py – містить класи, що визначають поля отриманих даних;
- project/middlewares.py – містить клас, що визначає операції, які мають виконуватись між отриманням відповіді сервера програмою Scrapy та передачі відповіді павуку;
- project/pipelines.py – дозволяє задати дії, що виконуються при відкриті/закриті павука (підпрограми для збирання даних);
- project/settings.py – містить налаштування користувача для павуків;
- project/spiders – директорія в якій зберігаються файли з класами павуків(збирачів даних).

При створені павука можна задати адреси, які будуть опитані, та яким чином будуть опрацьовані отримані данні. Це спрощує отримання даних для їх подальшої обробки комп'ютерною системою реєстрації новин на веб-ресурсах. Використання павука дозволить вказавши адреси новин, які представив користувачу інтернет-ресурс – одноманітно проаналізувати їх та зберегти результат в базу даних.

Звісно для початку нам потрібно отримати дані інтернет-ресурсу. Для цього ми можемо використати одну з двох базових бібліотек python – urllib [18] або urllib2 [19]. Обидві бібліотеки виконують URL-запити, проте мають різні функціональні можливості:

- для urllib2 можна вставити заголовки запиту шляхом створення об’єкту Request, а urllib може приямати як аргумент тільки URL;
- urllib дозволяє згенерувати рядки запиту GET за допомогою методу urlencode, а urllib2 такої функції не має. Це є одна з головних причин, чому варто використовувати urllib2 разом з urllib;

Проте можливі випадки, коли отримана відповідь сервера містить некоректні дані, або повідомлення про помилку. До цього може призвести одна з наступних проблем:

- агент користувача який був використано програмою було заблоковано веб-сервером;
- веб-сервер дозволяє доступ тільки з певних країн;
- веб-сторінка містить динамічний контент;
- веб-сторінка містить взаємодію з веб-формами;
- внутрішня помилка сервера;
- веб-ресурс використовує капчу (CAPTCHA).

В Python вже реалізовані засоби обходу більшості з цих проблем і тому достатньо написати всього кілька функцій для отримання даних навіть у випадку виникнення однієї з цих проблем.

Для того, щоб обійти блокування агента користувача, достатньо змінити його, передавши функції urllib2.Request() параметр «headers = {‘User-agent’: ‘MyAgent’}», де MyAgent назва агента користувача, що ймовірно не буде заблоковано сервером [19].

Для обходу блокування за країною, достатньо використати проксі-сервер з країни, де цей сайт буде доступній, urllib2 містить інструменти для реалізації підтримки проксі (proxy).

Для отримання динамічного контенту, що формується на веб-сторінці через скрипт мовою JavaScript, можна використати програмну бібліотеку управління браузерами Selenium WebDriver. За її допомогою можна отримати результат завантаження сторінки в браузері разом з динамічним контентом. Проте, ця програмна бібліотека витрачає багато ресурсів і тому її варто використовувати тільки, коли нам потрібно отримати динамічний контент сторінки [20].

Для отримання веб-сторінок, що потребують праці з веб-формами, можна використати модуль Mechanize, який не тільки робить поля форм легко доступними, але й бере на себе процес управління куками (cookies) веб-ресурсу.

Коли приходить відповідь, що свідчить про внутрішню помилку сервера (5xx), потрібно поставити запит в чергу очікування і повторити спробу пізніше.

Для опрацювання простої капчі (рис. 2.2) можна використати методи обробки зображень. Модуль Pillow реалізує клас Image, що надає методи обробки зображень [21]. За допомогою цього ми можемо спробувати прибрати фон зображення. Після цього ми можемо отримати текст за допомогою модуля Tesseract, що містить метод pytesseract.image_to_string(img) [22]. У випадку більш складної капчі (рис. 2.3), такий метод може не дати ніякого результата. Звічно, навіть в такому випадку ми можемо спробувати опрацювати таку капчу використавши один з онлайн-сервісів, які надають послуги в розгадуванні капчі реальними людьми через їх API, проте їх послуги можуть мати високу ціну.

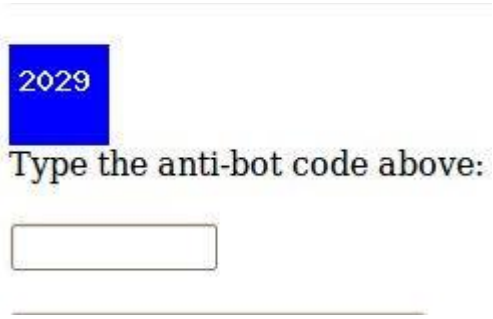

Рисунок 2.2 – Проста капча, що може бути розгадана простим переведенням зображення в рядок

Рисунок 2.3 – Складна капча, що може потребувати втручання людини для знаходження відповіді

Після того, як ми успішно отримали дані сторінки інтернет-ресурсу нам необхідно отримати вибірку новин, що знаходиться на цій сторінки. Для цього використовується BeautifulSoup, що дозволяє отримати DOM-структуру веб-сторінки. Це значно спрощує навігацію по HTML-документу та дозволяє користувачу явно задати, який блок містить вибірку інтернет-новин. Також перевагою BeautifulSoup є правильна інтерпретація структури веб-сторінки, навіть у випадку відсутності деяких з закриваючих тегів та лапків атрибутів, проте це інколи негативно впливає на швидкодію [23].

Отримавши вибірку новин, ми тепер можемо отримати їх адреси за допомогою регулярних виразів, який знайде всі веб-сторінки новин в

даному блоці. Після знаходження цих адресів ми можемо передати їх павуку Scrapy, який отримає та опрацює всі сторінки новин, що були представлені у цій вибірці.

Для зберігання інформації в базі даних було обрано нереляційну документо-орієнтовану систему керування базами даних з відкритим вихідним кодом із сімейства NoSQL, яка не потребує опису схеми таблиць – MongoDB [24].

MongoDB займає нішу між швидкими і масштабованими системами, що оперують даними у форматі ключ-значення, і реляційними СКБД, функціональними і зручними у формуванні запитів. В основі MongoDB лежить документо-орієнтована модель. На цьому варто додатково зосередити увагу. Використання документо-орієнтованої моделі означає, що вся інформація зберігатиметься у вигляді документів формату BSON, який походить від JSON (javascript object notation) [25]. BSON – це бінарний JSON, який дозволяє легко описувати об’єкти та інші структури даних і може бути прочитаний людиною. Документи в MongoDB зберігаються в колекціях. Якщо проводити паралель з реляційними базами даних (наприклад, MySQL [26]), то колекції – це таблиці, документи – кортежі, а стовпчики – атрибути.

Основні можливості MongoDB:

- документо-орієнтоване сховище (проста та потужна JSON-подібна схема даних);
- достатньо гнучка мова для формування запитів;
- динамічні запити;
- підтримка повноцінних індексів;
- профілювання запитів;
- простий та швидкий процес оновлення;

- ефективне зберігання бінарних даних великих обсягів, як фото та відео;
- логування операцій модифікування даних в БД;
- можливе функціонування відповідно до парадигми MapReduce [27].

Звісно у MongoDB є деякі недоліки. Наприклад, ця база даних не вміє самостійно зупиняти «повільні» запити, що негативно впливає на продуктивність системи, також вона має низьку продуктивність при виконані запиту count (рахування) на великих колекціях, проте ці недоліки не є критичними для комп'ютерної системи реєстрації новин на веб-ресурсах.

2.2. Метод навчання системи для розпізнавання деструктивного контенту

Для того, щоб почати навчання системи розпізнавання деструктивного контенту перш за все потрібно визначити, що таке деструктивний контент. Проблематика деструктивного контенту ще не достатньо вивчена і тому потребує додаткового аналізу.

Значення слова деструктивний, це направлений на руйнування, порушення функціонування чого-небудь [28]. У випадку глобальної мережі Інтернет за деструктивний контент приймемо наступне визначення, деструктивний контент, це контент, який містить інформацію, яка на пряму або опосередковано призводить до руйнування, порушення функціонування чого-небудь. Прикладами деструктивного контенту можуть бути повідомлення, що містять в собі агресію по відношенні до групи людей, заклики до насилля, державного перевороту, тощо. Також до

деструктивного контенту слід віднести повідомлення, що наносять моральної та ментальної шкоди читачу.

Можна виділити декілька методів боротьби з деструктивним контентом:

1) знищення деструктивного контенту – наприклад, використання блокувальників реклами для блокування реклами, що несе деструктивний характер;
2) оминання деструктивного контенту – наприклад, використання брандмауера для блокування сайтів, що користувач помітив як ті, що містять деструктивний контент;
3) розрідження деструктивного контенту – протидія деструктивному контенту в Інтернеті шляхом створення великої кількості корисного контенту на фоні якого деструктивний контент залишається не помітним.

В цій дипломній роботі було вирішено розвивати другий метод боротьби, шляхом створення комп'ютерної системи реєстрації новин на веб-ресурсах, яка має дати висновок про безпеку користування веб-ресурсом без втручання користувача, тим самим захищаючи його від навіть найменшої кількості шкоди, що він може зазнати під час оцінки веб-ресурса на безпеку у відношенні наявності деструктивного контенту.

Важливо визначити за якими саме критеріями текст новини можна віднести до деструктивного контенту. Через відсутність чітких критеріїв деструктивного контенту було створено ряд ознак, що текст містить деструктивний характер:

- текст містить агресивні висловлювання по відношенні однієї чи декількох соціальних, етнічних, національних або інших груп населення;

- текст містить образливі висловлювання по відношенню до почуттів певної групи населення (наприклад, образа почуттів віруючих);
- текст вводить в оману незахищених ментально людей (наприклад, дітей) і тим самим завдавати їм ментальної шкоди;
- текст створює фальшиву картину світу, тим самим «промиваючи» мізки незахищених ментально людей (наприклад, дітей);
- текст містить заклики до екстремізму, сепаратизму, тероризму, державного перевороту, тощо.

Підсумовуючи створенні ознаки деструктивного контенту, можна виділити три основні категорії деструктивного контенту:

- контент деструктивного характеру, що наносить шкоди особистості;
- контент деструктивного характеру, що наносить шкоди суспільству;
- контент деструктивного характеру, що має на меті задовільнити бажання третьої особи (екстремізм, сепаратизм, зміна державного ладу, тощо).

Для досягнення максимальної ефективності розпізнавання контенту деструктивного характеру, було створено 6 модулів перевірки тексту, до яких входить:

- модуль перевірки тексту на наявність контенту деструктивного характеру, що наносить шкоди особистості;
- модуль перевірки тексту на наявність контенту деструктивного характеру, що наносить шкоди суспільству;

- модуль перевірки тексту на наявність контенту деструктивного характеру, що наносить шкоди особистості або суспільству;
- модуль перевірки тексту на наявність контенту деструктивного характеру, що задовольняє бажання третьої особи;
- модуль перевірки тексту на наявність контенту деструктивного характеру, що задовольняє бажання третьої особи або наносить шкоди особистості;
- модуль перевірки тексту на наявність контенту деструктивного характеру, що задовольняє бажання третьої особи або наносить шкоди суспільству;
- модуль перевірки тексту на наявність контенту деструктивного характеру, що задовольняє бажання третьої особи або наносить шкоди особистості або суспільству.

Таке розділення на модулі дозволяє навчити систему таким чином, щоб вона могла розпізнавати навіть прихований деструктивний контент в тексті. Для навчання цих модулів, кожний з яких представляє з себе нейрону мережу, було підготовлено тільки 3 набори прикладів – по одному для кожної з категорій, які потім були передані модулям, що мають розпізнавати контент цієї категорії.

Як результат, запропонована система може не тільки розпізнати очевидну належність тексту новини до однієї з запропонованих категорій контенту деструктивного характеру, але й має можливість віднайти прихований деструктивний характер, завдяки модулям, що були навчені на наборах даних, що містили деструктивний контент різних категорій. Звісно, підвищення точності знаходження деструктивного контенту мало свою ціну – збільшення часу роботи алгоритму, проте перевага в точності проведення аналізу тексту набагато перевищує цей недолік.

2.3. Метод формування навчальних прикладів для системи розпізнавання деструктивного контенту

Метою формування прикладів для навчання системи розпізнавання деструктивного контенту є надати нейронній мережі такий набір вхідних даних, що дозволить їй в майбутньому ефективно розпізнавати наявність деструктивного контенту в тексті. Це ускладнюється тим, що система має знаходити не тільки явний деструктивний характер тексту, але й прихований. Для того, щоб задовільнити цю умову, необхідно створити такі навчальні приклади, що зможуть навчити систему розпізнавати деструктивний характер в цілому, а не тільки перевіряти його на вживані конструкції, адже тоді б не було потреби використовувати нейрону мережу.

Для того, щоб система могла розпізнати навіть прихований деструктивний контекст новин, створимо правила формування навчальних прикладів:

1) текст повинен містити контент деструктивного характеру всіх трьох категорій, при цьому відносний вміст контенту деструктивного характеру основної для даного тексту категорії повинен перевищувати вміст двох інших категорії, як мінімум в 4 рази;
2) текст повинен охоплювати більшість з можливих направлень дії деструктивного характеру, при цьому не повинно бути більше двох речень що містять однакове деструктивне направлення в одному тексті;
3) текст повинен містити від 100 до 1000 символів, при цьому вміст речень деструктивного характеру повинен бути в межах від 3 до 10%;

4) тексти повинні мати за собою якийсь сенс, а не просто набір речень;
5) теми текстів не повинні повторюватись;
6) навчальні приклади повинні охоплювати максимальну кількість категорій новин і при цьому кількість прикладів новин кожної категорії мають бути рівними.

Дотримуючись цих правил можна створити ідеальний набір тестових даних для навчання системи розпізнавання деструктивного контенту, що буде реагувати не просто на кількість слів-маркерів в тексті, але й знаходити підтекст новин, що тим самим дозволить нам виявити прихований деструктивний характер новини. Це значно збільшує точність з якою система після навчання знаходить деструктивний контент в тексті новин.

Звісно цей метод не надає абсолютної гарантії позитивного результату, проте за рахунок використання різних слів-маркерів і конструкцій деструктивного характеру – збільшується чутливість нейронної мережі, тим самим збільшуючи ймовірність успішного розпізнавання. На жаль, підвищення чутливість аналізу має і недолік – це збільшує кількість хибних спрацювань, що може зменшити достовірність аналізу. Проте головна мета розробки комп'ютерної системи реєстрації новин на веб-ресурсах є аналіз новин з зовнішніх ресурсів, що надає веб-ресурс, а тому збільшення похибки вимірювання в сторону частішого спрацювання не завдає значної шкоди, так як одночасно перевіряються декілька ресурсів. Результат цього навпаки є тільки позитивним, так як це дозоляє розкрити статті, що містять прихований деструктивний контент, навіть якщо ціна цьому незначне збільшення частоти хибного спрацювання.

2.4. Способи підготовки даних для навчання системи розпізнавання деструктивного контенту

На перший погляд для підготовки прикладів достатньо написати декілька сотень новин, що за змістом будуть відповідати розробленим критеріям деструктивного контенту. Проте система навчена таким чином буде мати не високу точність, так як стиль писання буде різним для автора тестових зразків та для реальних авторів новин деструктивного характеру. Найкращим варіантом буде, якщо кожна з новин буде написана різними авторами. Звісно можна взяти ці новини з різних сайтів, проте є ризик, що вибрані новини будуть мати одного автора. Зрозуміло, що це буде мати гарний ефект, так як ми зможемо з більшою точністю розпізнавати деструктивний контент написаний цим автором, проте це має таку саму ваду як і перший спосіб – нейроні мережі буде важче розпізнати деструктивний контент інших авторів.

Найкращим варіантом буде набрати контрольну групу авторів. Для підвищення ефективності бажано взяти людей різних професій та вікових груп. Для досягнення найкращого результату контрольна група має мати рівну кількість учасників кожної групи.

Для цього створимо наступні групи категоризації учасників контрольної групи. За віком учасників розділимо їх на:

- молодших учасників (10-15 років);
- молодих учасників (15-25 років);
- учасників середнього віку (25-35 років);
- досвідчених учасників (35-45 років);
- старших учасників (45+ років).

Учасників кожної вікової категорії має бути однакова кількість. В подальшому учасники в середині кожної науки за досвідом наукової праці діляться на:

- учасників, що зацікавлені в природничих науках;
- учасників, що зацікавлені в суспільних науках.

Звісно, учасників кожної з цих двох підгруп має бути рівна кількість в кожній віковій групі.

Наступним кроком, для того щоб охопити максимальну кількість можливих стилів написання, кожний учасник напише дев’ять новини, по три в кожній категорії деструктивного контенту. Дві з новин написаних учасниками в кожній категорії будуть написані на наукову і на суспільну теми з яскраво вираженим деструктивним характером. Третя і остання новина буде написана кожним учасником на вільну тему з завданням приховати деструктивний характер цієї новини.

Для досягнення найкращих результатів, контрольна група має складатися з 1000 учасників рівномірно розподілених за віком та напрямком зацікавленості в науці, що дасть нам 10 різних груп людей по 100 учасників.

Також покращити результат навчання може вибірка з справжніх новин деструктивного характеру, проте тут також бажано використати велику кількість людей для пошуку цих новин. Це дозволить охопити більшу кількість контенту і збільшить якість знайдених прикладів.

Для створення набору прикладів кращої якості, варто скомбінувати в рівних пропорціях кількість статей створених контрольною групою і знайдених на сторінках всесвітньої мережі [29].

2.5. Вхідні дані для опрацювання системою розпізнавання деструктивного контенту

Перед початком передачі даних нейронній мережі їх потрібно підготувати. Звісно можна просто передати весь текст новини, проте це буде мати мінімальну ефективність для нашої системи, якщо вона взагалі буде працювати. Для обробки текстових даних перед передачою їх нейронній мережі було обрано бібліотеку scikit-learn [30]. Вона має деталізовану документацію і реалізує більшість типових методів машинного навчання (МН). В цій бібліотеці реалізовані десятки алгоритмів для задач кластеризації, класифікації, методу опорних векторів, лінійної та логістичної регресії та десятків інших.

Scikit-learn побудована поверх SciPy (Scientific Python), який повинен бути встановлений перед використанням scikit-learn. Даний стек включає в себе:

- NumPy – розширення мови Python, що додає підтримку великих багатовимірних масивів і матриць, разом з великою бібліотекою високорівневих математичних функцій для операцій з цими масивами;
- SciPy – відкрита бібліотека високоякісних наукових інструментів для мови програмування Python;
- Matplotlib – бібліотека на мові програмування Python для візуалізації даних двовимірної графікою (3D графіка також підтримується);
- IPython – інтерактивна оболонка для мови програмування Python, яка надає розширену інтроспекцію, додатковий командний синтаксис, підсвічування коду і автоматичне доповнення;

- Sympy – бібліотека для символьних обчислень;
- Pandas – різні структури даних і аналіз.

Бібліотека scikit-learn надає реалізацію цілого ряду алгоритмів для навчання з учителем (Supervised Learning) і навчання без вчителя (Unsupervised Learning) через інтерфейс для мови програмування Python. Дана бібліотека поширюється під ліцензією «Simplified BSD License» і має дистрибутиви для безлічі різних версій Linux, заохочуючи тим самим академічне і комерційне використання scikit-learn. Ця бібліотека також допомагає вирішувати поставлені завдання в багатьох функціональних областей, до яких належать:

- Кластеризація (Clustering) – для групування нерозмічених даних, наприклад, метод k-середніх (k-means);
- Перехресна перевірка (Cross Validation) – для оцінки ефективності роботи моделі на незалежних даних;
- Набори даних (Datasets) – для тестових наборів даних і для генерації наборів даних з певними властивостями для дослідження поведінкових властивостей моделі;
- Скорочення розмірності (Dimensionality Reduction) – для зменшення кількості атрибутів для візуалізації та відбору ознак (Feature Selection), наприклад, метод головних компонент (Principal Component Analysis);
- Алгоритмічні композиції (Ensemble Methods) – для комбінування прогнозів декількох моделей;
- Витяг ознак (Feature Extraction) – визначення атрибутів в зображеннях і текстових даних;
- Відбір ознак (Feature Selection) – для виявлення значущих атрибутів на основі яких буде побудована модель;

- Оптимізація параметрів алгоритму (Parameter Tuning) – для отримання максимально ефективної віддачі від моделі;
- Множинне навчання (Manifold Learning) – для нелінійного скорочення розмірності даних;
- Алгоритми навчання з учителем (Supervised Models) – величезний набір методів не обмежується узагальненими лінійними моделями (Generalized Linear Models), дискримінантним аналізом (Discriminate Analysis), наївний баєсів класифікатор (Naive Bayes), нейронними мережами (Neural Networks), методом опорних векторів (Support Vector Machines) і деревами прийняття рішень (Decision Trees).

За допомогою scikit-learn ми проведемо токенізацію та індексацію тексту. Також ми можемо порахувати частотність слів в тексті та знайти слова-маркери даної новини.

Звісно, деструктивність контнету може бути досягнута не за рахунок вживаних слів, а за рахунок спеціального виділення слів або частин тексту. Для того, щоб не упустити важливість стилів та групування тексту, аналіз проводиться два рази – на основі всього тексту, та на основі субблоків в тексті. Тому нам потрібно провести токенізацію та індексацію, а також підрахунок частотності слів декілька раз – один раз для всього тексу і додатково для кожного виділеного блоку тексту. Це дозволяє досягти підвищення в точності оцінки тексту, так як береться до уваги не тільки параметри тексту в цілому, але й параметри кожної окремої частини тексту. Також до уваги беруться стилі цих виділених фрагментів тексту, що надає можливість більш точно оцінити текст на наявність контенту деструктивного характеру.

3. РОЗРОБКА ІНСТРУМЕНТАЛЬНИХ ЗАСОБІВ КОМП'ЮТЕРНОЇ СИСТЕМИ РЕЕСТРАЦІЇ НОВИН НА ВЕБ-РЕСУРСАХ

3.1. Алгоритм захоплення новин з веб-ресурсів

При створенні комп'ютерної системи реєстрації новин на веб-ресурсах однією з головних проблем є захоплення тексту новин з веб-ресурсів. Розроблений алгоритм (рис. 3.1) дозволяє не тільки обробити більшість проблем, що можуть виникнути при захопленні тексту веб-сторінки, але й зробити це ефективно, виконуючи мінімальну кількість операцій.

Після отримання адреси веб-ресурсу, що буде захоплено, система спочатку спробує отримати текст веб-сторінки, використовуючи тільки бібліотеки urllib та urllib2. В разі успішного отримання веб-сторінки, її вміст передається модулю реєстрації новин в базі даних.

У випадку, якщо сталась внутрішня помилка на сервері (програма отримала відповідь з кодом 5xx), то система тимчасово зупинить виконання програми, передавши управління на інші процеси системи, після цього програма спробує отримати текст програми в другий раз.

У випадку, якщо сервер відмовить в обслуговувані агенту користувача (User-Agent), що був використаний в запиті до сервера, то система змінить агент і спробує, ще раз.

У випадку, якщо сторінка містить динамічний контент, то буде використаний модуль Selenium WebDriver для завантаження сторінки.

У випадку, якщо сторінка містить капчу, що можна розгадати за допомогою бібліотек Pillow та Tesseract, то система повторить запит з результатом розпізнавання капчі.

Якщо нічого не допомогло,то система повідомить користувача про помилку.

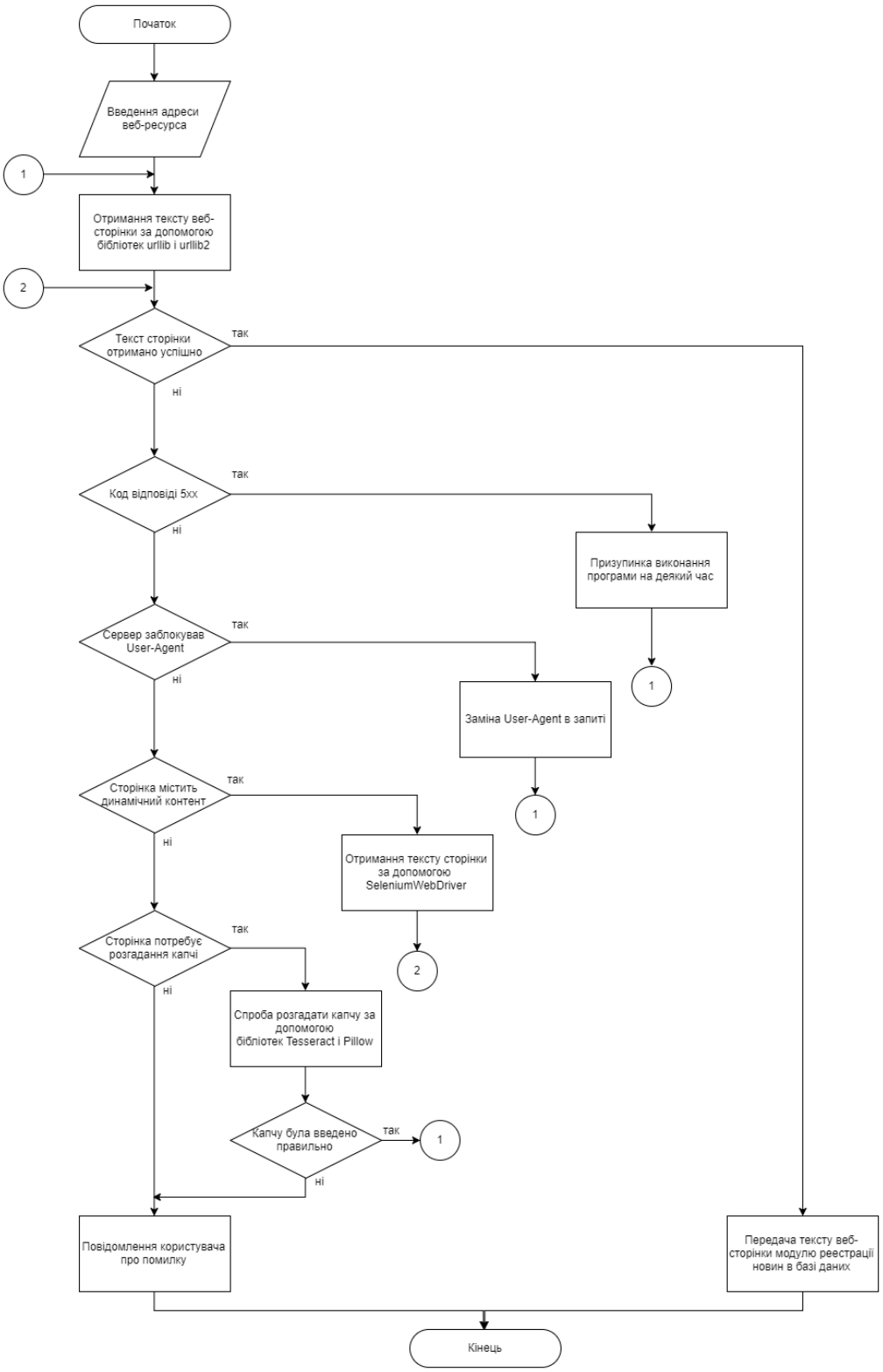

Рисунок 3.1 – Алгоритм захоплення новин з веб-ресурсів

3.2. Алгоритм реєстрації новин в базі даних

Не менш важливою проблемою є реєстрація новин в базі даних. Розроблений алгоритм (рис. 3.2) дозволяє не тільки обробити всі посилання на новини, які містить сторінка веб-ресурсу, але й уникнути повторного аналізу одного і того самого посиланні, або одного і того самого тексту новини.

Система із тексту веб-сторінки знаходить всі посилання на зовнішні джерела, після чого починає їх аналізувати.

Якщо в базі даних вже знаходиться результат аналізу новини з адресою отриманою з тексту веб-сторінки, то система пропускає це посилання і починає аналізувати наступне.

У випадку, якщо в базі не було знайдено даного посилання, то аналіз переходить до другого етапу – перевіряється чи вже знаходиться текст цієї новини в базі даних.

Якщо система знаходить новину в базі даних, то в запис бази даних про знайдену новину додається нове посилання, яке аналізувалось системою, а система переходить до аналізу наступного посилання.

Якщо під час двох перевірок новина не була знайдена, то система передає текст новини модулю аналізування. Результат аналізу і перетворений текст новини записуються в базу даних, після чого система переходить до аналізу наступного посилання.

Після того як система проаналізує всі знайдені посилання вона закінчує роботу модуля.

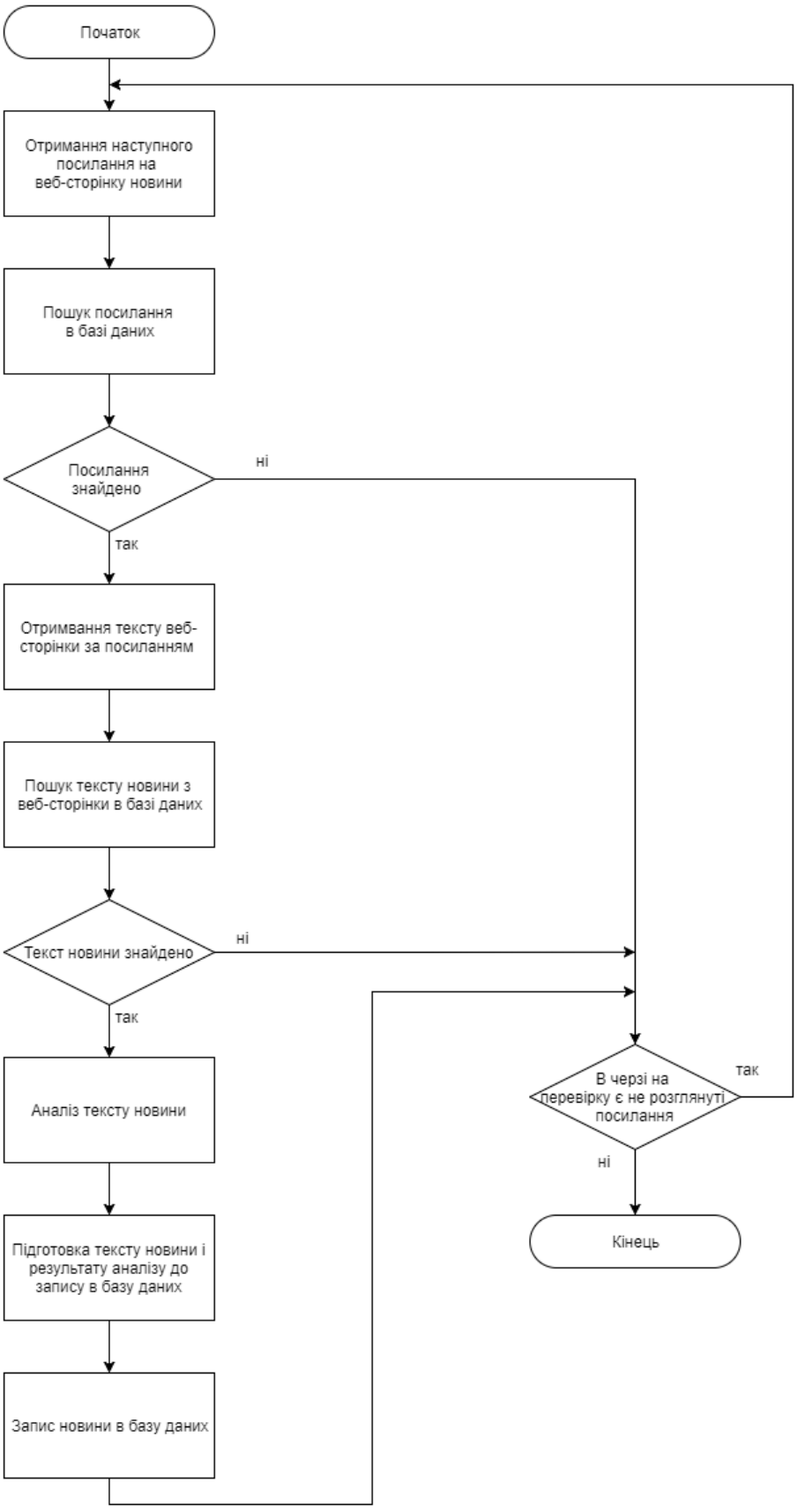

Рисунок 3.2 – Алгоритм реєстрації новин в базі даних

3.3. Алгоритм інтерпретації тексту новин комп'ютерною системою реєстрації новин на веб-ресурсах

Для проведення аналізу тексту на предмет присутності деструктивного контенту в тесті новини, необхідно спочатку провести частотний аналіз тексту, а також побудувати індекси. За це відповідає Модуль інтерпретації тексту, алгоритм роботи якого (рис 3.3) дозволяє проаналізувати частотність тексту та проіндексувати текст не тільки сторінки в цілому, але й кожний блок окремо. Результат виконання цього модуля можна передати нейронній мережі, що згенерує висновок про наявність деструктивного контенту в тексті.

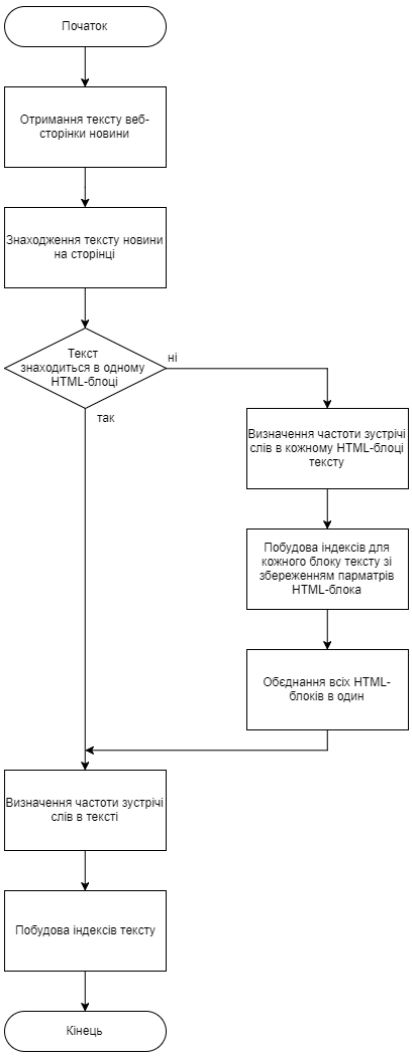

Рисунок 3.3 – Алгоритм інтерпретації тексту новин

3.4. Алгоритм роботи комп’ютерної системи реєстрації новин на веб-ресурсах

Алгоритм функціонування комп’ютерної системи реєстрації новин на веб-ресурсах (рис. 3.4) функціонує наступним чином.

1) За допомогою модуля захоплення новин (рис. 3.1) система отримує текст сторінки веб-ресурсу, що проходить перевірку.
2) Використовуючи регулярні вирази або прямі вказівки користувача, система знаходить блок сайту, що містить новини та посилання на відкритті джерела звідки вони були взяті.
3) В разі, якщо на другому кроці були знайдені посилання на відкриті джерела, то за допомогою модуля реєстрації новин (рис. 3.2) знаходяться та заносяться в базу даних новини за посиланням, а за допомогою модуля інтерпретації тексту новин (рис. 3.3) розраховується вміст деструктивного контенту за посиланнями.
4) За допомогою модуля реєстрації новин (рис. 3.2) заноситься в базу даних текст сторінки, а за допомогою модуля інтерпретації тексту новин (рис. 3.3) розраховується вміст деструктивного контенту для сторінки веб-ресурсу, що перевірявся.
5) Система повідомляє користувача про результат перевірки.

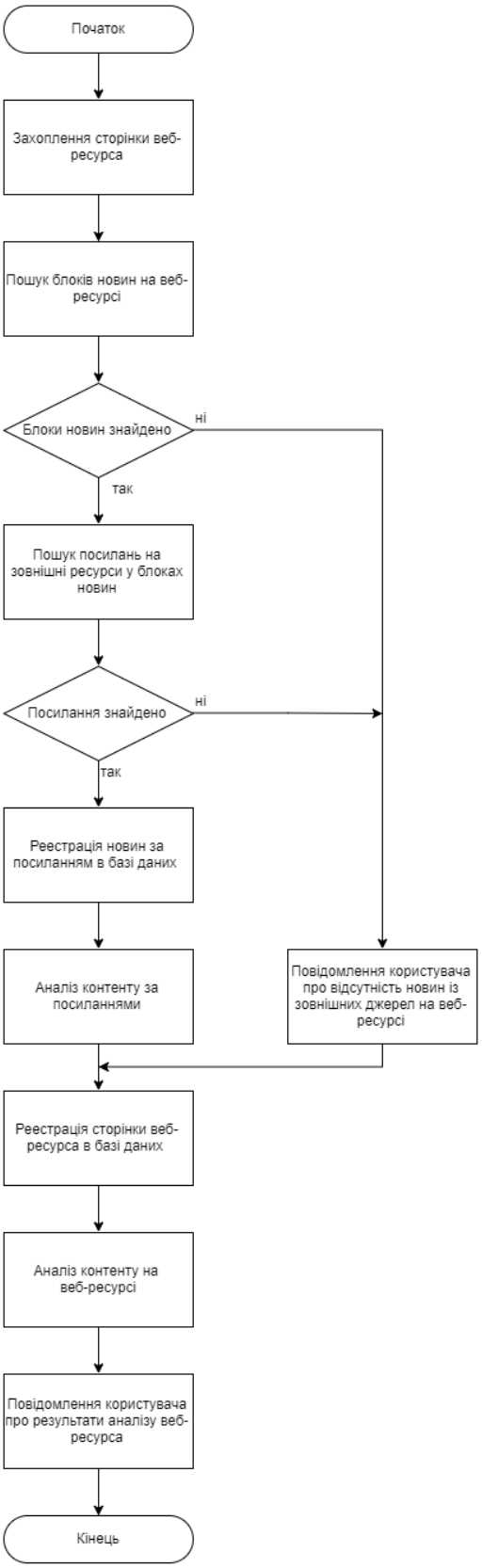

Рисунок 3.4 – Алгоритм роботи комп’ютерної системи реєстрації новин на веб-ресурсах

3.5. Структура комп’ютерної системи реєстрації новин на веб-ресурсах

Комп’ютерна мережа реєстрації новин на веб-ресурсах (рис. 3.5) складається з наступних модулів:

- модуль захоплення інформації – відповідає за отримання вихідного тексту сторінки веб-ресурсу в повному вигляді, не залежно від обмежень та механізмам протидії роботам веб-ресурса;
- модуль реєстрації інформації – відповідає за коректний запис інформації в базу даних, а також за пошук записів в базі даних за посиланням або текстом новини;
- модуль інтерпретації інформації – відповідає за пошук даних у вихідному коді сторінки веб-ресурсу та за пертворення тексту новини в вид готивий до передачі нейронній мережі комп’ютерної системи реєстрації новин на веб-ресурсах;
- модуль аналізу інформації – відповідає за виявлення деструктивного контенту на сторінці веб-ресурсу та за підготовку звіту про безпечність контенту веб-сторінки відносно наявності в ній потенційно небезпечного деструктивного контенту.

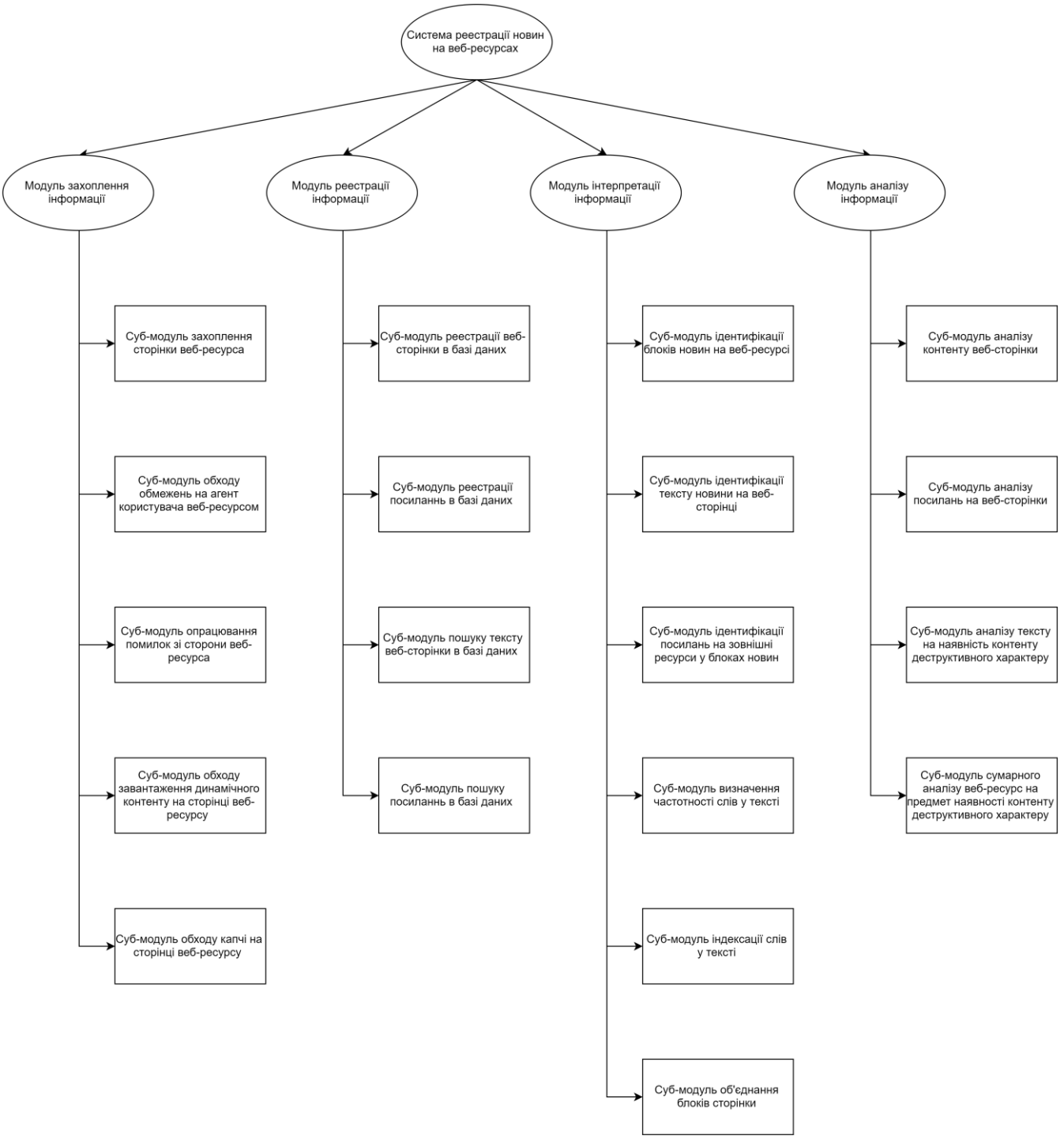

Рисунок 3.5 – Структура комп’ютерної системи реєстрації новин на веб-ресурсах

4. ОСОБЛИВОСТІ ТА ПЕРЕВАГИ РОЗРОБЛЕНОЇ СИСТЕМИ

4.1. Особливості комп’ютерної системи реєстрації новин на веб-ресурсах

Розроблена система реєстрації новин на веб-ресурсах має можливість розпізнавати наявність в тексті маркерів контенту деструктивного характеру, а також виносити попередження о підозрі про наявність в тексті прихованого деструктивного контенту.

Процес аналізу веб-ресурса комп’ютерною системою реєстрації новин на веб-ресрсах можна розділити на дев’ять кроків:

1. отримання вихідного тексту веб-сторінки;
2. знаходження блоку зовнішніх новин на веб-сторінці;
3. перевірки знаходження посилання на новину в базі даних;
4. якщо в третьому кроці не було знайдено запису в базі даних, перевірка знаходження тексту новини в базі даних;
5. якщо в третьому та четвертому кроці не було знайдено запису в базі даних, аналіз новини на наявність деструктивного контенту та занесення інформації в базу даних;
6. кроки 3-5 повторюються для кожної новини в блоці новин;
7. кроки 2-6 повторюються для кожного блоку новин на веб-сторінці;
8. кроки 1-7 повторюються кількість разі, яка буде достатньою для винесення висновку про безпеку веб-ресурса;
9. на основі аналізу готується висновок для користувача про безпеку відвідування веб-сторінки.

З результату аналізу користувач може дізнатись чи містить веб-ресурс новини деструктивного характеру, або новини, що підозрюються на

наявність прихованого деструктивного характеру. На основі цього користувач може вирішити чи безпечно буде відвідати цей ресурс чи можливо краще знайти альтернативу.

4.2. Тестування комп'ютерної системи реєстрації новин на веб-ресурсах

Тестування комп'ютерної системи реєстрації новин на веб-ресурсах встановило, що розроблена система відповідає всім заявленим вимогам. Завдяки використанню нейронних мереж система може розпізнати не тільки деструктивний підтекст новини завдяки словам-маркерам, але й винести підозру про наявність прихованого деструктивного контенту на базі схожості стилю написання новини до стилю написання новин, що містили деструктивний характер.

Проте, система все ще має простір до розвитку. Хоча система може чітко помічати новини з прихованим деструктивним контентом, проте це досягається за рахунок підвищеної чутливості, що призводить до спрацювання системи на деякі тексти, які не містять деструктивного контенту. Через це система не може винести однозначний висновок про наявність прихованого деструктивного контенту і тому може тільки виносити попередження користувачу про можливий вміст деструктивного контенту. Не зважаючи на це, система повноцінно виконує свої функції, і може майже однозначно встановити наявність не прихованого деструктивного контенту в тесті новини.

ВИСНОВКИ

Дипломна робота присвячена розв’язанню задачі розробки комп’ютерної системи реєстрації новин на веб-ресурсі. В процесі розв’язання отримано наступні результати.

1. Показано, що перспективним шляхом забезпечення ефективного розпізнавання контенту деструктивного характеру в новинах є використання сучасних типів нейронних мереж. Також визначено, що вдосконалення доцільно реалізувати за рахунок вдосконалення процесу навчання нейронних мереж.

2. Розроблена структура та алгоритмічне забезпечення комп’ютерної системи реєстрації новин на веб-ресурсах.

3. Розроблений процес генерування навчальних текстів для нейронної мережі, що пристосовано до розпізнавання в тексті не тільки яскраво вираженого деструктивного контенту, але й старанно прихованого.

4. Спроектовано програмне забезпечення комп’ютерної системи реєстрації новин на веб-ресурсах, що базується на використанні створеного алгоритмічного та інформаційного забезпечення.

5. В результаті проведеного аналізу показано доцільність застосування розроблених нейромережевих моделей для розпізнавання вмісту контенту деструктивного характеру в тексті новини.

6. Розроблену комп’ютерну систему рекомендується використовувати в складі інформаційних систем для підвищення рівня якості та безпеки користуванням веб-ресурсами користувачам всесвітньої мережі інтернет.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Наследов А.Д. Математические методы психологического исследования. Анализ и интерпретация данных : учебное пособие / А.Д. Наследов. – СПб. : Речь, 2004. – 392 с.
2. Деструктивний контент - повод для ограничения Интернета? [Електронний ресурс] – Режим доступу до ресурсу: https://rus.azattyq.org/a/destruktivniy-kontent-internet-ogranichenia/27739890.html.
3. УКР.НЕТ – головна сторінка [Електронний ресурс] – Режим доступу до ресурсу: https://ukr.net/ua/.
4. Google – головна сторінка [Електронний ресурс] – Режим доступу до ресурсу: https://www.google.com.ua/?hl=uk.
5. Яндекс – головна сторінка [Електронний ресурс] – Режим доступу до ресурсу: https://yandex.ua .
6. Smith J. Grow Your Business with Google AdWords: 7 Quick and Easy Secrets for Reaching More Customers with the World's #1 Search Engine / Jon Smith., 2009. – 144 с. – (1st Edition).
7. Elisan C. C. Malware, Rootkits & Botnets A Beginner's Guide/ Christopher C. Elisan., 2012. – 432 с. – (1st Edition).
8. Stewart J. M. Network Security, Firewalls And Vpns (Jones & Bartlett Learning Information Systems Security & Ass) (Standalone book) (Jones & Bartlett Learning Information Systems Security & Assurance) /J. Michael Stewar., 2013. – 490 с. – (2nd Edition).
9. uBlock Origin - An efficient blocker for Chromium and Firefox. Fast and lean. [Електронний ресурс] – Режим доступу до ресурсу: https://github.com/gorhill/uBlock.

10. Privacy Badger [Електронний ресурс] – Режим доступу до ресурсу: https://www.eff.org/privacybadger.
11. Ghostery [Електронний ресурс] – Режим доступу до ресурсу: https://www.ghostery.com/products/ .
12. AdGuard - The world's most advanced ad blocker! [Електронний ресурс] – Режим доступу до ресурсу: https://adguard.com/en/welcome.html.
13. AdBlock - You won't know what you're glad you're missing [Електронний ресурс] – Режим доступу до ресурсу: https://getadblock.com.
14. M. Benjes-Small C. The Library and Information Professional's Guide to Plug-ins and Other Web Browser Tools: Selection, Installation, Troubleshooting / C. M. Benjes-Small, M. L. Just., 2002.
15. Mitchell R. Malware, Web Scraping with Python: Collecting More Data from the Modern Web / Ryan Mitchell., 2018. – 308 с. – (2nd Edition).
16. Jacobson D. APIs: A Strategy Guide / D. Jacobson, G. Brail, D. Woods., 2011. – 150 с. – (1st Edition). – (Creating Channels with Application Programming Interfaces).
17. Scrapy [Електронний ресурс] – Режим доступу до ресурсу: https://scrapy.org.
18. urllib — URL handling modules [Електронний ресурс] – Режим доступу до ресурсу: https://docs.python.org/3/library/urllib.html.
19. urllib2 — extensible library for opening URLs [Електронний ресурс] – Режим доступу до ресурсу: https://docs.python.org/2/library/urllib2.html.
20. Selenium WebDriver [Електронний ресурс] – Режим доступу до ресурсу: https://www.seleniumhq.org/projects/webdriver/.

21. Pillow [Електронний ресурс] – Режим доступу до ресурсу: https://pillow.readthedocs.io/en/stable/.
22. Tesseract [Електронний ресурс] – Режим доступу до ресурсу: https://pypi.org/project/pytesseract/.
23. BeautifulSoup [Електронний ресурс] – Режим доступу до ресурсу: https://www.crummy.com/software/BeautifulSoup/.
24. Banker K. MongoDB in action / K. Banker. – NY : Manning, 2012. – 288 p.
25. Bassett L. Introduction to JavaScript Object Notation: A To-the-Point Guide to JSON / Lindsay Bassett., 2015. – 126 c. – (1st Edition).
26. MySQL [Електронний ресурс] – Режим доступу до ресурсу: https://www.mysql.com.
27. Kleppmann M. Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems / Martin Kleppmann., 2017. – 624 c. – (1st Edition).
28. Деструктивний [Електронний ресурс] – Режим доступу до ресурсу: http://ukrlit.org/slovnyk/деструктивний.
29. Aurélien G. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems / Aurélien Géron., 2017. – 574 c. – (1st Edition).
30. scikit-learn [Електронний ресурс] – Режим доступу до ресурсу: https://scikit-learn.org/stable/.